DOCUMENT RESUME

ED 439 145 TM 030 680

AUTHOR Lee, Guemin; Dunbar, Stephen B.; Frisbie, David A.
TITLE Measurement Models for a Testlet-Based Test.
PUB DATE 1999-04-22
NOTE 23p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Montreal, Quebec, Canada, April 19-23, 1999.)
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Factor Structure; *Goodness of Fit; *Measurement Techniques; *Models; Scores; Test Construction
IDENTIFIERS Congeneric Tests; *Testlets

ABSTRACT

It has been shown that the fundamental assumptions associated with conventional one-factor measurement models are frequently violated in analyses of scores from a test composed of testlets. Eight different measurement models were conceptualized for this kind of situation, and the goodness of fit of each model was examined. Measurement models incorporating correlated errors appear to be more appropriate than conventional measurement models with uncorrelated error specifications when testlets are involved. Also, the congeneric assumption about part tests or items appears to be more plausible than the essentially tau-equivalent assumption for a test composed of testlets. The one-factor congeneric model with correlated error specifications would be the best measurement model for a test composed of testlets if dichotomously-scored items are used as unit of analysis. However, a congeneric model using passage (testlet) scores can be considered as an alternative for a test composed of testlets when passage (testlet) scores are used as the unit of analysis. (Contains 6 tables and 19 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 439 145

Measurement Models for a Testlet-Based Test

Guemin Lee
CTB/McGraw-Hill

Stephen B. Dunbar
David A. Frisbie
University of Iowa

Paper Presented at the 1999 Annual Meeting
of the National Council on Measurement in Education
Montreal, Canada
April 22, 1999

TM030680

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
Guemin Lee
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

BEST COPY AVAILABLE

Abstract

It has been shown that the fundamental assumptions associated with conventional one-factor measurement models are frequently violated in analyses of scores from a test composed of testlets. Eight different measurement models were conceptualized for this kind of situation, and the goodness of fit of each model was examined. Measurement models incorporating correlated errors appear to be more appropriate than conventional measurement models with uncorrelated error specifications when testlets are involved. Also, the congenric assumption about part tests or items appears to be more plausible than the essentially tau-equivalent assumption for a test composed of testlets. The one-factor congeneric model with correlated error specifications would be the best measurement model for a test composed of testlets if dichotomously-scored items are used as unit of analysis. However, a congeneric model using passage (testlet) scores can be considered as an alternative for a test composed of testlet when passage (testlet) scores are used as the unit of analysis.

Measurement Models for a Testlet-Based Test

Testlets are small tests, small enough to manipulate but large enough to carry their own context (Wainer & Lewis, 1990; Wainer & Kiely, 1987). Previous studies have indicated that the reliability of testlet-based test scores is likely to be overestimated by conventional item-based reliability estimation methods (Sireci, Thissen & Wainer, 1991; Wainer & Thissen, 1996; Lee & Frisbie, in press). Wainer and Thissen (1996) and Sireci, Thissen and Wainer (1991) studied this topic using Bock's nominal model (Bock, 1972), a branch of item response theory, and concluded that the overestimation is due to "local dependence". Lee and Frisbie (in press), using a generalizability theory approach, provided reasons for the overestimation when coefficient alpha is used and contemplated the factors influencing the magnitude of the overestimation. However, none of this work resolves the question of which method or measurement model is most appropriate for estimating the reliability of test scores composed of testlets. The purposes of this study were to conceptualize the various kinds of measurement models for a test composed of testlets and to investigate the goodness of fit of those models to data. Because different measurement models have different assumptions, research on these models would provide empirical evidence about which assumptions are most essential to the application of the model.

In this paper, distinctions among eight measurement models are described in terms of the kind of parallelism exhibited by part tests or individual items: classically parallel, essentially tau-equivalent, and congeneric parts or items. The differences among these three classifications are due to differences in the distributions of observed scores, true scores, and error scores (Feldt & Brennan, 1989; Qualls, 1995). (In this paper, the classically parallel conceptualization will not be considered because it would be difficult to demonstrate, in a practical sense, that part tests or items of any test are truly parallel according to the classical model.)

The structural equation modeling (SEM) concept is another approach to defining measurement models. For each measurement model, a structural equation model connecting latent variables to one or more measures or observed variables is specified. That is, observed variables can be expressed as linear combinations of latent variables. The SEM measurement model represents the regression X on ξ and the elements of matrix Λ_X are the partial regression coefficients in the structural equations (Bollen, 1989; Jöreskog & Sörbom, 1993) :

$$X = \Lambda_X \xi + \delta \quad (1)$$

where X : observed variables
Λ_X : structural coefficients linking the latent and observed variables
ξ : latent variables
δ : error variables

With a set of structural equations, the covariance matrix of the observed variable vector X could be defined as Equation 2 (Bollen, 1989; Jöreskog & Sörbom, 1993a; Schumacker & Lomax, 1996):

$$\Sigma_X = \Lambda_X \Phi \Lambda_X' + \theta_\delta \quad (2)$$

where Σ_X : covariance matrix of observed variables
Λ_X : structural coefficients linking the latent and observed variables
Φ : covariance matrix of latent variables
θ_δ : covariance matrix of error variables

On the basis of the distinct conceptions about parallelism in part tests or items and with the SEM approach, the following eight measurement models were conceptualized for a test composed of testlets.

Model 1 : **One-Factor Essentially Tau-Equivalent Model**
In this model, all parameters in the Λ_X matrix are restricted to be equal, as the true score variances of items are assumed to be equal, and the θ_δ

matrix is restricted to be diagonal, which means the errors are not correlated, but the diagonal elements may differ, as the error variances of the items may differ.

Model 2 : **One-Factor Congeneric Model**
The elements of Λ_X may be different, which indicates a heterogeneity of true score variances among items, and θ_δ is restricted to the diagonal matrix like Model 1.

Model 3 : **One-Factor Essentially Tau-Equivalent Model with Correlated Errors**
The parameters in Λ_X are restricted to be equal. In the θ_δ matrix, the off-diagonal elements for within-passage items are allowed to have non-zero values, and the other off-diagonal elements, those for between-passage items, are restricted to have zero values.

Model 4 : **One-Factor Congeneric Model with Correlated Errors**
The matrix Λ_X is not restricted, and different parameters may be estimated. This model has the same specifications about the θ_δ matrix as Model 3.

Model 5 : **Multi-Factor Partially Tau-Equivalent Model**
In this model, testlet-specific common factors among items are considered. The within-passage items are assumed to be essentially tau-equivalent. However, the between-passage items may not be essentially tau-equivalent. Therefore, in order to describe this model, the label of "partially tau-equivalent model" is used. (Conceptually, the reliability estimates using this model should be the analog to stratified coefficient alpha.)

Model 6 : **Multi-Factor Congeneric Model**
This model also assumes testlet-specific common factors. However, the Λ_X matrix is not restricted for each testlet, and different parameters in each testlet may be estimated.

Model 7 : **One-Factor Essentially Tau-Equivalent Model using Testlet Scores**
This model has the same structure as Model 1 except that testlet (passage) scores are used instead of item scores.

Model 8 : **One-Factor Congeneric Model using Testlet Scores**
This model has the same structure as Model 2 but uses testlet (passage) scores rather than item scores.

The main purposes of this study were to conceptualize measurement models for a test composed of testlets and to assess the goodness of fit of those models to data. Three primary objectives of this study were to:

1. Examine the dimensionality and local dependence of tests composed of testlets to investigate the appropriateness of conventional unidimensional measurement models, based on dichotomously scored items, for describing test scores based on testlets.

2. Determine how well data from a test composed of testlets fit each of several measurement models.

3. Investigate the relationship between the degree of violation of the assumptions required by measurement modeling and the goodness of fit of measurement models to data.

Method

Data Source

The data for this study were taken from the 1995 Iowa Tests of Basic Skills (ITBS) Form M to Form K equating study. Grade 8 students were used. The tests to be used are the Reading Comprehension (Reading), Maps and Diagrams (Maps), and Math Problem Solving and Data Interpretation (Math) tests of the ITBS (Hoover, Hieronymous, Frisbie & Dunbar, 1994). These are testlet-based tests because each is composed of groups of items associated with its own stimulus material. The sample size and the general characteristics of each test are presented in Table 1.

Insert Table 1 About Here

A simulated data set was created to have the same structure as the grade 8 ITBS Vocabulary test. The simulated response data were generated by following the procedures used by Yen (1984), assuming item parameter estimates of the grade 8 Vocabulary test from 1992 the ITBS national standardization sample as the true item parameters. Though the simulated data set did not have naturally-formed testlets, seven testlets were randomly constructed for the purpose of comparison with tests composed of testlets.

Analyses

The local independence assumption was checked by Yen's (1984) Q_3 statistics. To examine the nature of conditional dependence measures, distributional characteristics of the pair of conditional dependence measures (one for within-passages and one for between-passages) were compared. The computer application program IRT_LD (Chen & Thissen, 1997) was used to compute Yen's Q_3 statistic. Two principal component analyses were completed to investigate issues related to the number of factors underlying a test. One set of analyses used a tetrachoric correlation matrix among individual items, and the other set used a product-moment correlation matrix among testlet scores. Sets of eigenvalues from each principal component analysis were compared.

For specified SEM measurement models for a test composed of testlets, the computer application program LISREL8 (Jöreskog & Sörbom, 1993a) was used to estimate model parameters and to compute the goodness of fit statistics. Various goodness of fit statistics for each model were compared among the eight measurement models to identify the most appropriate measurement models for conceptualizing a test composed of testlets.

Results

Local Independence

The distributional statistics for within-passage and between-passage Q_3 local item dependence measures are shown in Table 2. Even though the Q_3 statistic is a correlation between residuals of an item pair based on item response models (therefore, zero correlation might be expected for a locally independent item pair), Q_3 has a tendency to be slightly negative in the null case (Yen, 1984; Yen, 1993; Chen & Thissen, 1997). Yen (1993) demonstrated that the expected value of the Q_3 statistic, when local independence holds, is approximately $-1/(n-1)$, where n is the number of test items. The expected values for the Q_3

statistics, which are also presented in Table 2, can be used as a criterion for comparing the overall level of local dependence of within- and between-passage item pairs.

Insert Table 2 About Here

The averages of between-passage Q_3 statistics for the Reading, Maps, and Math tests have values similar to the expected values of Q_3 statistics, implying that item pairs between passages are locally independent. In contrast, the averages of within-passage Q_3 statistics for these tests have more positive values compared to the expected values of Q_3, even though the magnitudes of the differences in the Reading and Maps tests are greater than in the Math test. This finding suggests that the local item independence assumption is violated. For the simulated data set, because testlets were randomly constructed, averages of within- and between-passage Q_3 statistics are both similar to the expected value of Q_3.

Unidimensionality

Table 3 provides the first ten eigenvalues from tetrachoric correlation matrices based on individual items. These indicate that more than one factor would be required for explaining the data of the Reading and Maps tests. For the Math test and the simulated data set, one factor appears to be appropriate to explain the data.

Insert Table 3 About Here

To check the possibility of using passage (testlet) scores instead of using item scores, principal component analyses with product moment correlation matrices among testlet scores were conducted. Eigenvalues are presented in Table 4.

Insert Table 4 About Here

One dominant factor is evident, and the other eigenvalues are considered negligible. The well-known Kaiser (1970) criterion, retaining eigenvalues greater than unity, has been criticized

because of its susceptibility to the overidentification of dimensions (Cliff, 1988). Based on the Kaiser criterion, only one dimension is retained for all tests when testlet scores are used. In view of this susceptibility to overidentification, unidimensionality can be supported for the tests used in this study when testlet scores are used as the unit of analysis

Goodness of Fit Analysis

Schumacker and Lomax (1996) distinguished three types of goodness of fit statistics: model fit, model comparison, and model parsimony. Because the main focus of this study was on investigating the relative appropriateness of each measurement model to data, several goodness of fit statistics in the category of model fit were considered: χ^2 measure, root-mean-square residuals (RMR), and goodness of fit index (GFI). The χ^2 measure can be considered (N-1) times the minimum value of the fit function for the specified model, where N represents the number of examinees. Consequently, the χ^2 measure is sensitive to sample size, though it could be used as a statistical test. But, the analysis of the χ^2 measure in this study does not depend upon a statistical test; it depends instead upon a comparison of the magnitudes of the values obtained from various specified measurement models. RMR represents the square root of the mean squared differences between observed and model-based covariance matrixes. The smaller values of RMR represent better fit of the model to the data. GFI is based on a ratio of the sum of the squared difference between the observed and reproduced matrices to the observed variance. GFI values close to 1.0 reflect a good fit. The relative comparison of goodness of fit statistics among measurement models would be more meaningful than the absolute comparison between goodness of fit statistic and its criterion. These three kinds of fit statistics for each measurement model used in this study are presented in Table 5.

Insert Table 5 About Here

The essentially tau-equivalent model (Model 1) seems to present the worst fit compared to other measurement models, which are based on the same data set (Models 2 to 6) for the Reading, Maps, and Math tests. Among these measurement models based on dichotomous-scored items (Models 2 to 6), the one-factor congeneric model with correlated errors (Model 4) and the multi-factor congeneric model (Model 6) provide a relatively better fit to the data than do the other models for the Reading, Maps, and Math tests. For the simulated data set, three models within the essentially tau-equivalent family (Models 1, 3, and 5) seem similar, and three models within the congeneric family (Models 2, 4, and 6) provide goodness of fit statistics that are similar to one another.

Comparing the results from different content areas, the differences of fit statistics among measurement models within the same measurement family (either essentially tau-equivalent or congeneric) are greater than in the Reading and Maps tests than in the Math test. This finding can be explained in terms of the violation of the assumption for measurement modeling. That is, previously indicated, the assumptions for measurement modeling based on dichotomously-scored items are less violated in the Math test. So the less different fit statistics among measurement models within the same family in the Math test might not be so surprising. This relationship could be observed more evidently in the simulated data set. The simulated data set was included in this study for the purpose of comparison with tests composed of testlets. Consequently, it would be reasonable to expect little difference in fit statistics from different measurement models within the same family. The comparison of results from different contest areas having different degrees of violation of assumptions can be used as one piece of evidence to support the relationship between the degree of violation of assumptions and its effect on model-data fit.

Because Models 7 and 8 are based on passage (testlet) scores, it would not be reasonable to compare these models with other models (Models 1 to 6) based on dichotomously-scored items. Nonetheless, the appropriateness of these two models could be assessed by examining

fit indexes that are less sensitive to the scoring metric, such as RMR and GFI. Models 7 and 8 seem to be at least as appropriate as the other models in analyzing data from tests composed of testlets, when passage (testlet) scores instead of item scores are used as the unit of analysis. The congeneric model using testlet scores (Model 8) represents a better fit than the essentially tau-equivalent model using testlet scores (Model 7) except for the Maps test. It seems reasonable to expect that, if passage (testlet) scores instead of item scores were used, the true score variances for passage (testlet) scores be different one another in a test because numbers of items within passages are different from passage to passage.

The remaining part is devoted to investigate the specific issues related to tests composed of testlets: uncorrelated vs. correlated error, one-factor vs. multi-factor, and essentially tau-equivalent vs. congeneric issues. These issues can be addressed by using the difference of the χ^2 measures from two different measurement models. These differences of the χ^2 measures and their statistical test results are presented in Table 6.

Insert Table 6 About Here

For the Reading, Maps, and Math tests, the difference of the χ^2 measures between one-factor essentially tau-equivalent model (Model 1) and one-factor essentially tau-equivalent model with correlated errors (Model 3), and difference between one-factor congeneric model (Model 2) and one-factor congeneric model with correlated errors (Model 4) are very large compared to their degrees of freedom. This means that the uncorrelated error assumption required by classical test theory or the local independence assumption required by item response theory is not satisfied in tests composed of testlets. Because item responses for the simulated data set were generated under the local independence assumption and the testlets in this data set were constructed randomly, the uncorrelated error assumption (or local independence assumption) seems to be satisfied in this case. Similar results, shown in Table 6,

were obtained by comparing the one-factor essentially tau-equivalent model (Model 1) with the multi-factor partially tau-equivalent model (Model 5) and the one-factor congeneric model (Model 2) with the multi-factor congeneric model (Model 6).

For the case of essentially tau-equivalent vs. congeneric models, congeneric models provide a better fit to data than do essentially tau-equivalent models for every test used in this study. However, this is more evident in comparing models based on dichotomously-scored items than models based on testlet scores.

Discussion

When items in a test are related to the same single passage or other stimulus material, there might be statistical dependence among those items. Items within a certain testlet share something in common even after eliminating the influence of the general common factor from every item. Consequently, a fundamental assumption required by measurement modeling (the uncorrelated error assumption for classical test theory or the local independence assumption for item response theory) is frequently violated in a test composed of testlets.

Measurement models incorporating correlated errors or multi-factors appear to be more appropriate than conventional one-factor measurement models with uncorrelated error specifications when testlets are involved. Also, the congenric assumption about part tests or items appears to be more plausible than the essentially tau-equivalent assumption for a test composed of testlets, at least when dichotomously-scored items are used as the unit of analysis.

One important consideration should be addressed at this point: the relative appropriateness of the multi-factor vs. correlated error models. This question cannot be answered by a statistical approach, but rather it should be investigated by a conceptual analysis of the test structure. For example, if the testlet factor is conceptualized as fixed, it would reasonable to treat testlets with multi-factor measurement models. In contrast, if the

testlet factor is considered random, one-factor measurement models with a correlated error specification would be better.

Whether a factor is random or fixed in a particular situation depends on the sampling plan used to form the test. For the situation in this study, the sampling plan usually assumed a universe of passages, from which several passages were randomly sampled. Items in a passage were also assumed to be sampled randomly for that passage. Then, the passages are considered random in this sampling plan (Lee & Frisbie, in press). Conceptually, one-factor measurement models with correlated errors would be more appropriate for this case than multi-factor measurement models might be.

The statistical appearances of tests composed of various testlets (random or fixed) would be similar if the individual items were taken as the fundamental measurement unit (e.g. there would be a systematic pattern in the covariance matrix among individual items). For this reason, applying measurement models based on a structural equation modeling approach could not make clear distinctions between measurement models with correlated errors and multi-factor measurement models. However, because different ways of treating testlets (random or fixed) could lead to a dramatic difference in measurement applications, the sampling plan for a test should be considered before applying a specific measurement model.

In conclusion, when items are used as the unit of analysis, the assumptions required by measurement modeling for tests composed of testlets are violated to some degree, but those assumptions are satisfied when using passage (testlet) scores as the unit of analysis. The one-factor essentially tau-equivalent model (the same for Cronbach's alpha) presents the worst model-data fit for a test composed of testlets. The one-factor congeneric model with correlated error specifications would be the best measurement model among the six models based on dichotomously-scored items that were conceptualized in this study. A congeneric model using passage (testlet) scores can be considered as an alternative for a test composed of testlet, as long as passage (testlet) scores are used as the unit of analysis. An increase in the extent of

violation of the assumptions required by measurement modeling leads to a corresponding increase in difference of fit statistics between uncorrelated and correlated measurement models, and between one-factor and multi-factor measurement models.

References

Bock, R.D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*(1), 29-51.

Bollen, K.A. (1989). *Structural equations with latent variables.* NY: John Wiley & Sons.

Chen, W.-H., & Thissen, D. (1997). Local *dependence indexes for item response theory. Journal* of Educational and Behavioral Statistics, 22, 265-289.

Cliff, N. (1988). The eigenvalues-greater-than-one rule and the reliability of components. *Psychological Bulletin, 103,* 276-279.

Feldt, L.S., & Brennan, R.L. (1989). Reliability. In R.L. Linn (Ed.) *Educational measurement.* Washington, DC: American Council on Education.

Fleishman, J., & Benson, J. (1987). Using LISREL to evaluate measurement models and scale reliability. *Educational and Psychological Measurementt, 47,* 925-938.

Hoover, H.D., Hieronymus, A.N., Frisbie, D.A., & Dunbar, S.B. (1994) *Iowa Tests of Basic Skills : Interpretive guide for school administrators.* Chicago, IL: The Riverside Publishing Company.

Jöreskog, K.G., & Sörbom, D. (1993a). *LISREL8 User's reference guide.* Chicago, IL: Scientific Software International, Inc.

Jöreskog, K.G., & Sörbom, D. (1993b). *PRELIS2 User's reference guide.* Chicago, IL: Scientific Software International, Inc.

Kaiser, H.F. (1970). A second generation little jiffy. *Psychomerika, 35,* 401-415.

Lee, G., & Frisbie, D.A. (in press). Estimating reliability under a generalizability theory model for test scores composed of testlets. *Applied Measurement in Education.*

Schumacker, R.E., & Lomax, R.G. (1996). *A beginner's guide to structural equation modeling.* Mahwah, NJ: Lawrence Erlbaum Associates.

Sireci, S.G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement. 28*(3), 237-247.

Qualls, A.L. (1995). Estimating the reliability of a test containing multiple item formats. *Applied Measurement in Education , 8*(2), 111-120.

Wainer, H., & Kiely, G.L. (1987). Item clusters and computerized adaptive testing : A case for testlets. *Journal of Educational Measurement, 24*(3), 185-201.

Wainer, H., & Lewis, C. (1990). Toward a psychometrics for testlets. *Journal of Educational Measurement, 27*(1), 1-14.

Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measurement : Issues and Practice, 15*(1), 22-29.

Yen, W.M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8,* 125-145.

Yen, W.M. (1993). Scaling performance assessments: strategies for managing local item dependence. *Journal of Educational Measurement, 30,* 187-213.

TABLE 1
Descriptive Statistics for Data Sources Used in This Study

Characteristic	Reading	Maps	Math	Simulation
Sample Size	663	632	537	1000
No. of Items	49	33	36	43
No. of Passages	8	5	8	7
No. of Items per Passage	9,4,7,5,5,6,3,10	7,7,6,6,7	8,4,4,4,4,4,4,4	7,6,6,6,6,6,6
$\bar{X}$	24.9	16.3	16.5	26.4
S_X	9.99	6.34	6.38	8.47
Skewness	0.375	0.383	0.363	-0.196
Kurtosis	2.216	2.306	2.413	2.280

Note. Reading = Reading Comprehension, Maps = Maps and Diagrams, Math = Math Problem Solving and Data Interpretation, Simulation = Simulated data.

TABLE 2

Distribution of Yen's Q_3 Statistics for Within-Passage and Between-Passage Item Pairs

Test	No. of Q_3	E (Q_3)	Mean	Diff.	S.D.	Range
Reading	1176	-.021				
Between	1030		-.021	.000	.043	-.180~.121
Within	146		.038	.059	.058	-.131~.196
Maps	528	-.031				
Between	435		-.029	.002	.045	-.177~.098
Within	93		.031	.062	.049	-.064~.164
Math	630	-.029				
Between	560		-.022	.007	.049	-.183~.137
Within	70		.008	.037	.057	-.087~.166
Simulation	903	-.024				
Between	792		-.019	.005	.034	-.151~.107
Within	222		-.016	.008	.031	-.088~.052

Note. Reading = Reading Comprehension, Maps = Maps and Diagrams, Math = Math Problem Solving and Data Interpretation, Simulation = Simulated data; E (Q_3) = Expected value of Q_3, Diff. = Absolute value of the difference between E (Q_3) and the sample mean.

TABLE 3

First Ten Eigenvalues of the Tetrachoric Correlation Matrices Based on Individual Item Scores

Eig	Reading		Maps		Math		Simulation	
Rank	Eig	Diff	Eig	Diff	Eig	Diff	Eig	Diff
1	14.07	11.15	7.84	6.00	8.41	6.42	12.87	11.45
2	2.92	1.23	1.84	0.34	1.99	0.19	1.42	0.09
3	1.68	0.17	1.50	0.07	1.80	0.34	1.32	0.06
4	1.51	0.19	1.43	0.11	1.46	0.11	1.27	0.06
5	1.32	0.01	1.32	0.08	1.35	0.07	1.20	0.06
6	1.31	0.06	1.24	0.05	1.28	0.05	1.14	0.05
7	1.25	0.05	1.19	0.04	1.23	0.04	1.09	0.03
8	1.20	0.02	1.15	0.09	1.19	0.05	1.06	0.00
9	1.18	0.03	1.06	0.01	1.14	0.07	1.06	0.02
10	1.15	0.03	1.05	0.05	1.07	0.02	1.03	0.01

Note. Reading = Reading Comprehension, Maps = Maps and Diagrams, Math = Math Problem Solving and Data Interpretation, Simulation = Simulated data, Eig = eigenvalue, and Diff = difference between consecutive eigenvalues.

TABLE 4
Eigenvalues of the Product-Moment Correlation Matrices Based on Passage Scores

Eig	Reading		Maps		Math		Simulation	
Rank	Eig	Diff	Eig	Diff	Eig	Diff	Eig	Diff
1	4.10	3.30	2.59	1.91	3.34	2.51	4.26	3.68
2	0.80	0.17	0.68	0.06	0.83	0.01	0.58	0.09
3	0.63	0.03	0.62	0.05	0.82	0.10	0.49	0.03
4	0.60	0.07	0.57	0.03	0.72	0.05	0.46	0.01
5	0.53	0.07	0.54		0.67	0.06	0.45	0.04
6	0.46	0.01			0.61	0.04	0.41	0.06
7	0.45	0.04			0.57	0.12	0.35	
8	0.41				0.45			

Note. Reading = Reading Comprehension, Maps = Maps and Diagrams, Math = Math Problem Solving and Data Interpretation, Simulation = Simulated data, Eig = eigenvalue, and Diff = difference between consecutive eigenvalues.

TABLE 5
Goodness of Fit Statistics of Specified Measurement Models

GoF	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
Reading Comprehension Test								
χ^2	2328.0 df=1175	2089.3 df=1127	1494.5 df=1029	1282.1 df=981	1734.3 df=1140	1518.1 df=1099	96.9 df=27	70.8 df=20
RMR	0.067	0.046	0.058	0.036	0.057	0.038	0.065	0.032
GFI	0.85	0.86	0.91	0.93	0.90	0.91	0.96	0.97
Maps and Diagrams Test								
χ^2	992.5 df=527	817.5 df=495	613.4 df=435	473.9 df=403	759.9 df=513	610.1 df=485	9.7 df=9	8.1 df=5
RMR	0.064	0.042	0.053	0.031	0.057	0.036	0.025	0.018
GFI	0.91	0.92	0.95	0.96	0.93	095	0.99	0.99
Math Problem Solving and Data Interpretation Test								
χ^2	1125.7 df=629	811.2 df=594	932.5 df=559	641.5 df=524	1004.0 df=594	704.5 df=566	70.5 df=27	36.1 df=20
RMR	0.078	0.042	0.073	0.038	0.073	0.040	0.071	0.028
GFI	0.88	0.92	0.91	0.94	0.90	0.93	0.97	0.98
Simulated Data								
χ^2	1499.8 df=902	1113.2 df=860	1369.3 df=792	983.9 df=750	1428.7 df=875	1095.1 df=839	47.5 df=20	25.1 df=14
RMR	0.061	0.029	0.060	0.027	0.058	0.029	0.047	0.013
GFI	0.93	0.95	0.94	0.96	0.93	0.95	0.99	0.99

Note. GoF = Goodness of fit statistic, χ^2 = Chi-square measure, RMR = Root-mean-square residual, GFI = Goodness-of-fit index.

TABLE 6

Difference of χ^2 Statistics from Two Compared Measurement Models and Statistical Tests

Test	Compared Models	χ^2 difference	df difference	Probability
	Uncorrelated vs. Correlated Error Models			
Reading	1 vs 3	833.5	146	.00000
	2 vs. 4	807.2	146	.00000
Maps	1 vs. 3	379.1	92	.00000
	2 vs. 4	343.6	92	.00000
Math	1 vs. 3	193.2	70	.00000
	2 vs. 4	169.7	70	.00000
Simulation	1 vs. 3	130.5	110	.08870
	2 vs. 4	129.3	110	.10091
	One-Factor vs. Multi-Factor Models			
Reading	1 vs. 5	593.7	35	.00000
	2 vs. 6	571.2	28	.00000
Maps	1 vs. 5	232.6	14	.00000
	2 vs. 6	207.4	10	.00000
Math	1 vs. 5	121.7	35	.00000
	2 vs. 6	106.7	28	.00000
Simulation	1 vs. 5	71.1	27	.00008
	2 vs. 6	18.1	21	.64267
	Essentially Tau-Equivalent vs. Congeneric Models			
Reading	1 vs. 2	238.7	48	.00000
	3 vs. 4	209.4	48	.00000
	5 vs. 6	216.2	41	.00000
	7 vs. 8	26.1	7	.00048
Maps	1 vs. 2	175.0	32	.00000
	3 vs. 4	139.5	32	.00000
	5 vs. 6	149.8	28	.00000
	7 vs. 8	1.6	4	.80879
Math	1 vs. 2	314.5	35	.00000
	3 vs. 4	291.0	35	.00000
	5 vs. 6	299.5	28	.00000
	7 vs. 8	34.5	7	.00001
Simulation	1 vs. 2	386.6	42	.00000
	3 vs. 4	385.4	42	.00000
	5 vs. 6	333.6	36	.00000
	7 vs. 8	22.4	6	.00102

Note. Reading = Reading Comprehension, Maps = Maps and Diagrams, Math = Math Problem Solving and Data Interpretation, Simulation = Simulated data, χ^2 difference = difference of χ^2 measures from two compared measurement models, df difference = difference of degrees of freedom from two compared measurement models.